CALIBRATE WHAT YOU TRUST: RELIABLE MULTIMODAL 3D PANOPTIC SEGMENTATION UNDER DOMAIN SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

We present the first study of Domain Generalization (DG) for multimodal 3D panoptic segmentation (mm-3DPS), targeting robust perception under unseen domain shifts in autonomous driving. We identify two key challenges: reliable multimodal representation learning when individual modalities are affected differently by domain shifts, and reliable panoptic prediction when these shifts propagate to the final scene decomposition. To address them, we propose **RAMP3D**, a DG framework specifically designed for mm-3DPS. RAMP3D incorporates Reliability-guided Multimodal Adaptation (ReMA), which explicitly estimates modality reliability and adaptively regulates multimodal adaptation under domain shift. It further introduces Reference-based Calibration (ReCal), which leverages stable scene-relative cues to separately calibrate stuff-region predictions and thing-instance scores. Extensive experiments across diverse shifts in time, weather, location, and sensor conditions show that RAMP3D consistently outperforms strong baselines, including adaptations of state-of-the-art 3D DG and multimodal UDA methods.

1 INTRODUCTION

Autonomous driving and robotic perception require a structured understanding of both individual objects and the surrounding environment Feng et al. (2026); Zhai et al. (2026). By assigning semantic categories to all points while distinguishing individual *thing* instances from *stuff* regions, 3D panoptic segmentation unifies semantic and instance-level scene understanding. Recent multimodal methods further enhance panoptic perception by fusing LiDAR geometry with image semantics as a complementary feature Pan et al. (2025); Li et al. (2026); Mohan et al. (2026). However, multimodal complementarity depends on the reliability of modality-specific cues. Under domain shift, a mismatch may emerge between what the model trusts and what remains reliable, affecting both multimodal representation and panoptic prediction. We study these failures under the domain generalization (DG) setting, where training uses only labeled source data and requires robust generalization to multiple unseen target domains without access to target-domain data.

First, maintaining effective multimodal representation learning becomes challenging when the reliability of image and LiDAR cues changes across domains. Cross-modal consistency Jaritz et al. (2020); Cardace et al. (2023) and VFM priors Wu et al. (2024); Li et al. (2025) can improve knowledge transfer, but these strategies alone do not ensure reliable representation when modality-specific cues deteriorate. Visual cues can deteriorate under changes in illumination or weather, and even strong VFMs remain constrained by the quality of their input images. Multimodal models trained on source data may still over-rely on these degraded cues, limiting generalization to unseen domains. Table 1 illustrates this limitation: multimodal IAL and its VFM-based variants underperform LiDAR-only IAL on the nighttime target. The substantial source-to-nighttime performance drop of image-only models further highlights the vulnerability of image-based predictions under this shift. Together, these observations motivate multimodal representation learning that explicitly accounts for modality reliability.

Second, stuff and thing predictions can fail in different ways under domain shift. Existing multimodal DG methods primarily target semantic segmentation Li et al. (2025) or object detection Wu et al. (2026), leaving the distinct reliability requirements of stuff-region assignment and thing-instance

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

selection in panoptic segmentation insufficiently addressed. For stuff classes, adverse weather such as rain can degrade both image and LiDAR cues, making spatial assignments ambiguous. Fig. 1(a) illustrates poorly delineated boundaries between stuff masks. For thing instances, selecting candidates based on semantic classification confidence can retain inaccurate instance masks while suppressing valid candidates under domain shift. As shown in Fig. 1(b), high-confidence candidates with little overlap with any ground-truth instance can be retained (red crosses), whereas class-correct candidates with high-quality masks can be excluded from the final prediction (green points).

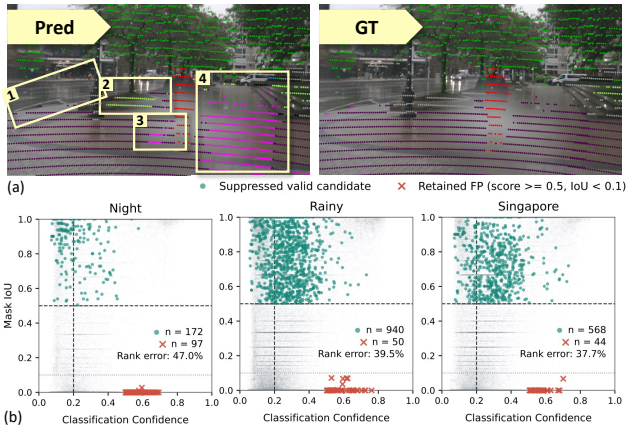


Figure 1: (a) Comparison of predictions and ground truth, highlighting segmentation errors of stuff classes under rainy conditions. (b) Failure cases of confidence-based thing-candidate selection. Gray points denote all thing proposals. Green dots highlight class-correct, high-IoU proposals corresponding to missed ground-truth instances. Red crosses highlight retained high-confidence false positives with little ground-truth overlap.

To address these challenges, we propose **Reliability-Aware Multimodal 3D Panoptic Segmentation (RAMP3D)**, the first framework for domain-generalized multimodal 3D panoptic segmentation, which improves reliability at both the *feature* and *panoptic prediction* levels. At the feature level, we introduce a reliability-guided multimodal adaptation (**ReMA**) module to explicitly estimate modality reliability. ReMA employs a set of learnable tokens as cross-modal shared references and leverages *2D semantic evidence*, *3D geometric evidence*, and their *cross-modal agreement* to estimate regional reliability for each modality. The resulting reliability dynamically modulates multimodal feature adaptation, suppressing degraded visual cues while preserving informative image semantics, thereby improving the generalizability of multimodal representations.

At the panoptic level, we propose a reference-based calibration (**ReCal**) mechanism to address the distinct uncertainties in stuff-region assignment and thing-instance selection. For stuff classes, where domain shift can blur spatial assignments and boundaries, ReCal exploits relative appearance and road-referenced geometric cues to correct voxel-wise mask predictions. For thing classes, where prediction confidence may become misaligned with actual instance quality, ReCal combines relative features with semantic and mask consistency to estimate instance quality and recalibrate candidate scores.

Our main contributions are summarized as follows.

- We are the first work to study domain generalization for multimodal 3D panoptic segmentation and introduce **RAMP3D** to improve reliability at both the representation and prediction levels.
- We propose **ReMA**, which estimates modality reliability to regulate image- and LiDAR-derived residual updates within a frozen VFM.
- We develop **ReCal**, which uses a shared within-scene road reference to calibrate stuff-region assignment and re-score thing candidates.
- Extensive experiments across intra-dataset and cross-dataset shifts demonstrate that RAMP3D consistently outperforms strong baselines, including LiDAR-based DG and multimodal UDA methods adapted to the DG setting.

2 Related Work

Multimodal 3D Panoptic Segmentation. LiDAR panoptic segmentation has evolved from detection-driven Sirohi et al. (2021); Xu et al. (2023); Ye et al. (2023) and clustering-based pipelines Zhou

et al. (2021); Hong et al. (2021); Razani et al. (2021); Xu et al. (2022); Li et al. (2022) to query-based frameworks that directly predict thing and stuff masks (Gu et al., 2022; Marcuzzi et al., 2023; Xiao et al., 2025), yet the performance remains limited due to the sparsity of LiDAR points. Multimodal methods therefore complement LiDAR with dense image semantics through fusion strategies (Zhang et al., 2023; Song et al., 2024; Yu et al., 2025), integration at the data, feature, and query levels (Pan et al., 2025), or pretrained visual-semantic priors (Li et al., 2026). UP-Fuse (Mohan et al., 2026) suppresses uncertain camera cues under sensor degradation. However, all these methods are trained and evaluated within a closed domain. Cross-domain 3D panoptic segmentation has mainly been studied under unsupervised domain adaptation (UDA), such as the LiDAR-only method AdaptLPS (Besic et al., 2022) and multimodel method PanDA (Pan et al., 2026); both of which require access to target-domain data. We instead study this task under the DG setting, where no target data are available.

Domain Generalization for Semantic Segmentation. Domain generalized semantic segmentation (DGSS) learns from labeled source domains to perform well on unseen target domains. Early 2D methods suppress domain-specific feature statistics (Choi et al., 2021), while recent methods preserve the knowledge of vision foundation models (VFMs) through parameter-efficient adaptation. For example, REIN (Wei et al., 2024) refines frozen VFM features with learnable tokens, GeCo (Zang et al., 2026) regularizes such adaptation with geometry-consistent perturbations, and DepthForge (Chen et al., 2025) injects depth cues estimated by a frozen depth VFM into learnable tokens. In 3D, LiDAR-only methods simulate sparsity variations (Kim et al., 2023) or learn sensor-robust features through bird's-eye-view auxiliary supervision (Saltori et al., 2023), and multimodal methods build on cross-modal learning (Jaritz et al., 2020) with BEV-based fusion (Li et al., 2023), VFM prompting (Wu et al., 2024), or diffusion priors (Li et al., 2025). These multimodal methods mainly learn transferable representations, but do not estimate whether each modality remains reliable under different target domains.

Reliability-Aware Multimodal Fusion and Prediction. Under domain shift, reliability issues arise at both the feature level, where corrupted modalities can be over-trusted during fusion (Yang et al., 2024), and the prediction level, where confidence scores may no longer reflect prediction quality. At the feature level, MetaBEV (Ge et al., 2023) and MoME (Park et al., 2025) handle sensor failures via selective BEV aggregation and expert routing, respectively, and Cocoon (Cho et al., 2025) weights modality features by object-level uncertainty. However, these methods assume that training and test data share the same distribution. For domain generalization (DG), CCF (Wu et al., 2026) addresses imbalanced modality utilization in multimodal 3D detection without explicitly estimating modality reliability. In contrast, our ReMA (Sec. 3.2) explicitly estimates the region-wise reliability of both modalities, and uses it to gate the adaptation of a frozen VFM. At the prediction level, Mask Scoring R-CNN (Huang et al., 2019) and 3D instance segmentation methods (Vu et al., 2022; Sun et al., 2023) re-score instances by predicted mask IoU. Our ReCal (Sec. 3.3) extends such calibration to panoptic outputs using a road reference computed from the model's own predictions in each frame, re-scoring thing candidates with road-relative and semantic-consistency cues and correcting stuff mask logits with road-relative features and local height information.

3 Methodology

3.1 Task Formulation and Overview

Task Formulation. We study single-source domain generalization (DG) for multimodal 3D panoptic segmentation (mm-3DPS). The model is trained on a *single* labeled source domain $\mathcal{S}$ and evaluated on *multiple* unseen target domains $\{\mathcal{T}\}$, without accessing target-domain data during training. At test time, given synchronized RGB images $\mathbf{x}^{\text{2D}}$ and a LiDAR point cloud $\mathbf{x}^{\text{3D}}$ from an unseen target domain, the model predicts a 3D panoptic segmentation $\hat{\mathbf{y}}$ that assigns a semantic label to every point and an instance identity to points belonging to *thing* classes. Here, *thing* classes denote countable objects with distinct instances, whereas *stuff* classes describe amorphous regions without instance-level distinctions.

Framework Overview. We propose a Reliability-Aware Multimodal 3D Panoptic Segmentation method, named **RAMP3D**, illustrated in Fig. 2. We adopt IAL Pan et al. (2025) as the base mm-3DPS architecture, with a frozen DINOv2 Oquab et al. (2024) image encoder and a range-view FRNet Xu

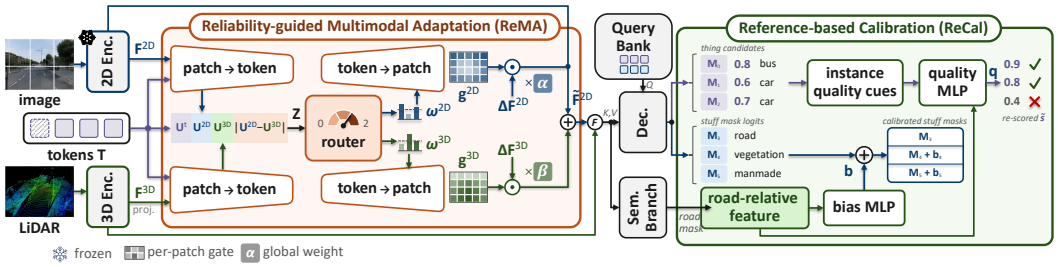


Figure 2: **Framework overview of our proposed RAMP3D**. RAMP3D improves the robustness of multi-modal 3D panoptic segmentation under domain shifts from two complementary perspectives: *reliability-aware multimodal representation* and *reference-based panoptic calibration*. **ReMA** (left) aggregates image and LiDAR evidence through shared learnable tokens to estimate modality reliability. Local reliability gates and learnable global weights regulate modality-conditioned residual updates within a frozen VFM. The adapted image features are then fused with LiDAR features for panoptic decoding. **ReCal** (right) uses a shared within-scene road reference to calibrate stuff mask logits and re-score thing candidates according to their estimated mask quality, improving prediction reliability under domain shift.

et al. (2025) LiDAR encoder. Following IAL, image features are lifted to 3D space and fused with LiDAR features, followed by panoptic query initialization and transformer decoding.

RAMP3D improves robustness under domain shift at two complementary stages. At the representation level, ***reliability-guided multimodal adaptation*** (Sec. 3.2) estimates cross-modal reliability and uses modality-specific, per-patch gates to regulate 2D- and 3D-conditioned residual updates before multimodal feature fusion. At the prediction level, ***reference-based calibration*** (Sec. 3.3) uses the road region as a shared intra-scene reference to calibrate the two components of panoptic prediction: stuff mask logits and thing-instance class predictions.

3.2 Reliability-guided Multimodal Adaptation (ReMA)

ReMA builds upon REIN, which introduces a parameter-efficient approach for adapting a frozen vision foundation model (VFM) using a small set of learnable tokens that inject instance-aware residual refinements into intermediate features, thereby preserving pretrained knowledge while adapting the representation for domain-generalized segmentation. ReMA extends this mechanism to the *multimodal setting* by incorporating LiDAR-derived residual updates and explicitly estimating modality reliability to regulate the contributions of image- and LiDAR-derived information during VFM adaptation. ReMA comprises four steps: First, patch-to-token attention pools image and LiDAR features into compact token-wise representations. Second, a router uses these representations and their cross-modal discrepancy as evidence to predict modality-specific reliability weights. Third, the same attention maps project these weights back to patches, yielding per-patch reliability gates. Finally, The per-patch gates, together with learnable global scalars, regulate the modality-conditioned residual updates used for VFM adaptation.

Patch-to-Token Attention. At each layer, we introduce a set of learnable tokens $\mathbf{T} \in \mathbb{R}^{(M+1)\times C}$ to serve as shared references across the image and LiDAR modalities. Following REIN Wei et al. (2024), the first token contributes no content: it participates in attention normalization but is excluded from aggregation sum. We omit the layer index for simplicity. $\mathbf{T}$ is parameterized through a low-rank factorization to reduce the number of trainable parameters. To aggregate patch-level features into compact token-wise evidence, we first establish patch-to-token associations. Let $\mathbf{F}^{2\mathrm{D}}, \mathbf{F}^{3\mathrm{D}} \in \mathbb{R}^{N\times C}$ denote the image features and the spatially aligned LiDAR features of one view, where N is the number of patches. We build $\mathbf{F}^{3\mathrm{D}}$ by projecting the point features into each camera, averaging those that fall into the same patch. Patches receiving no point stay zero. The modality-specific attention maps are computed as

$$\mathbf{S}^k = \mathrm{Softmax}\left(\frac{\mathbf{F}^k\mathbf{T}^\top}{\sqrt{C}}\right), \qquad k \in \{2\mathrm{D}, 3\mathrm{D}\}, \tag{1}$$

where $\mathbf{S}^k \in \mathbb{R}^{N\times(M+1)}$ contains the attention weights from patches to shared tokens.

Evidence-conditioned Routing. Using these patch-to-token associations, we aggregate image and LiDAR features into paired token-wise evidence for reliability-aware routing. Let $\hat{\mathbf{S}}^k$ denote $\mathbf{S}^k$ with the empty-token column removed and each remaining column normalized to sum to one. We then construct the modality-specific evidence as:

$$\mathbf{U}^{\mathrm{k}} = \phi^{\mathrm{k}}\Big((\hat{\mathbf{S}}^{\mathrm{k}})^\top \mathbf{F}^{\mathrm{k}}\Big). \tag{2}$$

where ϕ^{k} denotes a linear projection followed by LayerNorm, which compresses the $C = 768$-dimensional evidence into a $D = 32$-dimensional space. The same projection applied to the non-empty tokens gives the token-level evidence $\mathbf{U}^t \in \mathbb{R}^{M\times D}$. We concatenate the three sources together with their cross-modal discrepancy to form the router input:

$$\mathbf{Z} = \left[\mathbf{U}^t, \mathbf{U}^{2\mathrm{D}}, \mathbf{U}^{3\mathrm{D}}, \left|\mathbf{U}^{2\mathrm{D}} - \mathbf{U}^{3\mathrm{D}}\right|\right], \tag{3}$$

and predict a reliability weight for each modality:

$$\boldsymbol{\omega}^k = 2\sigma\big(\psi^k(\mathbf{Z})\big), \tag{4}$$

where ψ is a two-layer MLP with GELU that outputs two channels. The reliability weight $\boldsymbol{\omega}^k \in (0, 2)$, where values below, equal to, and above 1 correspond to suppression, neutral weighting, and enhancement. The router votes on the two modalities independently, without constraining them to sum to one, allowing both modalities to be enhanced when both are reliable or suppressed when neither is reliable. We zero-initialize the final layer of ψ, so training starts at $\boldsymbol{\omega} = \mathbf{1}$.

Token-to-Patch Gating. We map the token-wise reliability weights back to patches as follows:

$$\mathbf{g}^k = \mathbf{S}^k \left[1,\ (\boldsymbol{\omega}^k)^\top\right]^\top \tag{5}$$

where the scalar 1 assigns a fixed neutral weight to the empty token. Since each row of $\mathbf{S}^k$ sums to one, the gate values are attention-weighted averages of the token-wise weights and therefore remain in $(0, 2)$. This mapping reuses the attention maps, so per-patch reliability needs no extra dense prediction head.

Reliability-guided VFM Adaptation. We use the same mechanism as REIN Wei et al. (2024) to compute the interaction between learnable tokens and modality features as per-modality residual updates $\Delta\mathbf{F}^{2\mathrm{D}}$ and get $\Delta\mathbf{F}^{3\mathrm{D}}$ in the same way. We scale each update twice, once globally and once per patch,

$$\tilde{\mathbf{F}}^{2\mathrm{D}} = \mathbf{F}^{2\mathrm{D}} + \alpha\left(\mathbf{g}^{2\mathrm{D}} \odot \Delta\mathbf{F}^{2\mathrm{D}}\right) + \beta\left(\mathbf{g}^{3\mathrm{D}} \odot \Delta\mathbf{F}^{3\mathrm{D}}\right), \tag{6}$$

where $\odot$ broadcasts each gate over the channel dimension. The learnable scalars α and β carry the overall importance of the two branches, while $\mathbf{g}^{2\mathrm{D}}$ and $\mathbf{g}^{3\mathrm{D}}$ vary with the frame and the patch. The updated features $\tilde{\mathbf{F}}^{2\mathrm{D}}$ enter the next transformer block.

Augmentation and Supervision. Following common practice in domain generalization, we apply region-wise dropout to both modalities during training, exposing the router to varying local reliability. Dropout is applied independently to each modality with probability 0.5, with dropped regions defined by GridMask Wu et al. (2026). We further apply semantic supervision to the learnable tokens. Semantic logits are computed as dot products between voxel features and token embeddings and supervised by the cross-entropy and Lovász losses as $\mathcal{L}_{\mathrm{rema}}$.

3.3 Reference-based Calibration

Reliable multimodal representations alone do not guarantee reliable panoptic predictions, as the decoder's mappings from features to masks and confidence scores may not generalize beyond the source domain. We introduce **Reference-based Calibration (ReCal)**, which uses a shared within-frame road reference for two branch-specific corrections: adjust stuff mask logits of stuff masks and quality estimation for thing candidate re-scoring.

Shared Road Reference. Road regions typically provide broad spatial support and a consistent semantic role in driving scenes. We use the base model's auxiliary semantic branch to identify predicted road voxels and average their fused multimodal features to obtain a frame-specific reference $\boldsymbol{\mu}_{\mathrm{road}}$. For each voxel, subtracting this reference from its fused feature $\mathbf{f}_{\mathrm{fuse}}$ yields the road-relative feature $\mathbf{r}$:

$$\mathbf{r} = \mathbf{f}_{\mathrm{fuse}} - \boldsymbol{\mu}_{\mathrm{road}}. \tag{7}$$

This within-frame subtraction can reduce feature offsets shared across the scene. If no road voxels are predicted, we use the mean fused feature over all voxels in the frame as the reference.

Stuff Region Calibration. We combine road-relative features with geometric height cues to modulate voxel-wise stuff mask responses. Using a local elevation-map representation Steinke et al. (2024), we aggregate the heights of predicted road voxels within BEV neighborhoods to estimate the local road height h at each voxel's horizontal position. We concatenate the road-relative feature $\mathbf{r}$ and height $z - h$ to predict a spatial bias:

$$b = \phi_b([\mathbf{r};\, z - h])\,, \tag{8}$$

where z is the voxel's height coordinate and ϕ_b is a two-layer MLP with GELU activation. We set b to zero at voxels predicted as road. At each voxel, the bias is added to all stuff mask logits. Let M denote a stuff mask logit at a given voxel and decoder stage. The calibrated logit is $\tilde{M} = M + b$. The voxel-wise biases are computed once per frame and reused for the initial mask prediction and all subsequent decoder layers. Calibrated intermediate masks guide feature aggregation for the next query update, while the final calibrated masks are used to generate the final panoptic mask.

Thing Quality Estimation and Re-scoring. For each thing query, we obtain its predicted region by thresholding the final mask probabilities. We average-pool the road-relative voxel features within this region to obtain a regional road-relative descriptor $\bar{\mathbf{r}}$. To characterize candidate quality, we extract four *instance quality cues*. The vector $\boldsymbol{\pi}_{\text{sem}} \in \mathbb{R}^{N_{\text{cls}}}$ is the average semantic distribution within the predicted region, where N_{cls} is the number of classes. The scalar η_{rank} is the percentile rank of the query's classification confidence among all thing queries in the frame, with larger values indicating higher confidence. The scalar η_{mask} is the mean mask probability within the predicted region. Finally, η_{purity} is one minus the normalized entropy of the regional semantic distribution, with larger values indicating greater semantic concentration. We concatenate these statistics with the road-relative feature within corresponding regions and predict a quality score:

$$q = \sigma(\phi_q([\bar{\mathbf{r}}; \boldsymbol{\pi}_{\text{sem}}; \eta_{\text{rank}}; \eta_{\text{mask}}; \eta_{\text{purity}}]))\,, \tag{9}$$

where ϕ_q is a three-layer MLP and σ is the sigmoid function. We combine the original classification confidence s with q through a weighted geometric mean:

$$\tilde{s} = s^{\,1-\gamma} q^{\,\gamma}, \tag{10}$$

where γ controls the contribution of the quality score. The calibrated score $\tilde{s}$ is used for candidate filtering and voxel assignment during panoptic fusion. The thing re-scoring branch adjusts candidate scores while leaving the predicted classes and mask logits unchanged. We set $\gamma = 0.5$ following the default setting Sun et al. (2023).

Training Objective. Inspired by prior work on IoU-based instance scoring Huang et al. (2019); Sun et al. (2023), we derive quality supervision from mask IoU. For each candidate, we compute the maximum IoU between its predicted region and the ground-truth thing instances in the frame. Candidates with maximum IoU above 0.5 form the positive set $\mathcal{P}$, while those below 0.1 form the negative set $\mathcal{N}$. The remaining candidates are excluded from quality supervision. We apply a pairwise ranking loss to encourage positive candidates to receive higher quality logits than negative candidates:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{P}||\mathcal{N}|} \sum_{p \in \mathcal{P}} \sum_{n \in \mathcal{N}} \log\left(1 + \exp(\xi_n - \xi_p)\right), \tag{11}$$

where ξ denotes the pre-sigmoid quality logits, i.e., the inputs to $\sigma(\cdot)$ in Eq. 9, and p and n represent positive and negative candidates, respectively. We additionally apply binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$ to the quality scores, assigning targets of 1 to positive candidates and 0 to negative candidates. Both quality losses are computed only for frames containing positive and negative candidates and are averaged over these valid frames.

The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{rema}} + (\mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{BCE}})\,, \tag{12}$$

where $\lambda = 0.2$, $\mathcal{L}_{\text{base}}$ is the base model's panoptic loss and $\mathcal{L}_{\text{rema}}$ is the semantic loss defined in Sec. 3.2.

4 EXPERIMENTAL RESULTS

4.1 EXPERIMENTAL SETUP

We evaluate domain generalization under both intra-dataset and cross-dataset shifts. Models are trained on labeled source-domain data and evaluated on unseen target domains.

Intra-Dataset Domain Shifts. We construct an intra-dataset domain generalization benchmark on nuScenes Caesar et al. (2020); Fong et al. (2022), which provides data collected using a 32-beam LiDAR and six surround-view RGB cameras. We define domain shifts based on three attributes in the nuScenes metadata: geographic location, weather, and illumination. The source domain consists of samples collected in Boston under sunny daytime conditions, which are regarded as relatively clean conditions. Each target is selected by its dominant attribute while the remaining attributes are left unconstrained: *Singapore* for geographic shift, *rainy* for weather shift, and *nighttime* for illumination shift. This reflects the realistic deployment, where a primary condition change rarely occurs in isolation. Following standard mm-3DPS protocols Pan et al. (2026; 2025); Zhang et al. (2023), we retain LiDAR points across the full sensing range, use all six camera views, and map the annotations to 10 *thing* and 6 *stuff* classes.

Cross-Dataset Domain Shift. For cross-dataset generalization, we use the full nuScenes training set as the source domain and evaluate on SemanticKITTI Behley et al. (2019; 2021) and Waymo Sun et al. (2020). We choose nuScenes as the source dataset because it provides more complete LiDAR-camera coverage. SemanticKITTI is collected in Germany using a 64-beam LiDAR and two forward-facing cameras, while Waymo is collected in the United States using five LiDARs and five front- and side-facing cameras. These datasets introduce larger shifts in sensor characteristics and scene distributions compared with the intra-dataset setting. For Waymo, we construct panoptic annotations by combining the official semantic and instance segmentation annotations. Following prior work Jaritz et al. (2022); Wu et al. (2024); Pan et al. (2026), we map the dataset-specific annotations to a shared label space containing 1 *thing* and 5 *stuff* classes. Evaluation is performed on the full point clouds for all datasets. Detailed dataset configurations, label mappings, and dataset statistics are provided in the appendix.

Implementation Details. We build our framework upon IAL Pan et al. (2025), replacing its 3D backbone with FRNet Xu et al. (2025). The range-view resolution is set to 32×960. For the image branch, we replace the original backbone with a pretrained DINOv2 ViT-B/14 and resize images from all camera views to 640×360. We freeze the DINOv2 parameters and adapt its intermediate features through ReMA at each transformer layer. Following REIN Wei et al. (2024), we employ a pixel decoder to generate hourglass-style multi-scale features and use 100 learnable tokens. All models are trained for 80 epochs using the same data augmentation and optimization settings as IAL. The base learning rate is 2×10^{-4}, with a $4\times$ learning-rate multiplier for the 3D encoder. Training is conducted on four NVIDIA A40 or RTX 5090 GPUs, with a batch size of 2 per GPU.

Evaluation Metrics. We use standard panoptic segmentation metrics for evaluation. The primary metric is panoptic quality (PQ), defined as the product of segmentation quality (SQ) and recognition quality (RQ).

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\mathrm{TP}|}}_{\mathrm{SQ}} \times \underbrace{\frac{|\mathrm{TP}|}{|\mathrm{TP}| + \frac{1}{2}|\mathrm{FP}| + \frac{1}{2}|\mathrm{FN}|}}_{\mathrm{RQ}}. \tag{13}$$

4.2 COMPARED METHODS

As no prior work targets domain-generalized multimodal 3D panoptic segmentation, we construct LiDAR-only, image-only, and multimodal baselines within the IAL Pan et al. (2025) panoptic segmentation framework. All variants use the same panoptic decoder, training schedule, and base data augmentations. Variants using LiDAR features share the same 3D backbone, while image-backbone replacements are described below. The use of VFMs is indicated in the second column of Table 1. $\mathrm{IAL}_{3\mathrm{D}}$ and $\mathrm{IAL}_{2\mathrm{D}}$ are constructed by training IAL with only LiDAR or image features, respectively. Because camera views do not cover all LiDAR points, image-only variants, marked with *, are evaluated only on image-covered points.

We incorporate four representative methods into these baselines: *REIN* Wei et al. (2024) is proposed for 2D semantic segmentation, and we apply it to the image branch of both $\mathrm{IAL}_{2\mathrm{D}}$ and multimodal IAL. It replaces the original image encoder with a frozen DINOv2 ViT-B/14 and adapts its intermediate

Table 1: Domain generalization results under intra-nuScenes and cross-dataset settings. Italicized headers indicate sources; others are target sets. *US S.D.* denotes samples collected from Boston under sunny daytime; sKITTI denotes SemanticKITTI. Row colors distinguish LiDAR-only, image-only, and multimodal methods. * indicates evaluation only on image-covered LiDAR points. † indicates using frozen VFM without adaptation. All results are reported as PQ (%).

		Intra-nuScenes				Cross-dataset		
Method	**VFM**	*US S.D.*	**Nighttime**	**Rainy**	**Singapore**	*nuScenes*	**sKITTI**	**Waymo**
IAL_{3D}	-	74.99	46.98	68.30	57.07	80.99	40.10	35.95
+ DCGL	-	74.61	42.96	60.06	66.90	80.53	42.04	28.10
IAL^{*}_{2D}	-	63.78	6.40	53.57	41.44	71.49	36.78	24.47
+ REIN*	DINOv2	75.24	22.34	66.41	49.37	76.27	36.48	28.99
IAL	-	76.25	41.62	68.75	60.24	79.63	25.85	20.15
+ VFM†	DINOv2	77.38	42.08	71.99	64.69	80.70	35.90	26.39
+ REIN	DINOv2	79.48	44.01	74.51	66.41	80.89	32.61	29.76
+ DCGL	-	78.15	43.74	72.90	62.66	80.01	40.77	22.81
+ xMUDA	-	78.38	42.46	71.30	61.36	80.66	39.15	30.29
+ UniDSeg	CLIP	79.01	44.14	73.29	63.09	80.43	40.75	32.27
RAMP3D	DINOv2	**80.12**	**52.06**	**77.65**	**69.02**	**82.18**	**46.13**	**43.80**

features through learnable tokens. These variants evaluate the effect of adapting the image branch with a VFM. *DCGL* He et al. (2025) learns category-level geometric representations and encourages their consistency between original and perturbed point clouds. We incorporate this geometric learning strategy into the LiDAR branch of both IAL_{3D} and multimodal IAL. *xMUDA* Jaritz et al. (2022) introduces cross-modal consistency between image and LiDAR. Building on this cross-modal learning framework, *UniDSeg* Wu et al. (2024) replaces the image encoder with a frozen CLIP ViT-B and incorporates its designed modules to adapt visual features using geometric information.

4.3 Main Results

Table 1 shows that RAMP3D achieves the best PQ on every source and target domain in both settings.

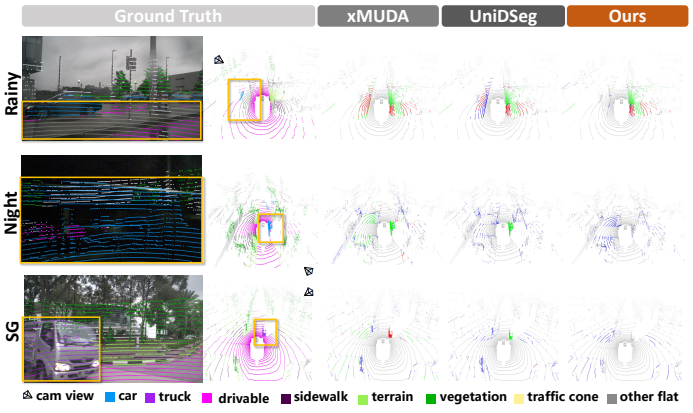


Figure 3: Visualization of panoptic segmentation error maps. We present camera images with semantic-colored GT for reference.

Intra-dataset Domain Generalization. RAMP3D consistently improves generalization across all three intra-nuScenes target domains. Compared with the multimodal IAL model, it improves PQ by 10.44, 8.90, and 8.78 points on nighttime, rainy, and Singapore, respectively. The nighttime results highlight that incorporating image features does not guarantee improved generalization. Multimodal IAL falls 5.36 PQ points below its LiDAR-only counterpart. Adapting the image branch with DINOv2 and REIN reduces this gap, but still underperforms the LiDAR-only baseline by 2.97 PQ points. RAMP3D exceeds IAL+REIN and LiDAR-only IAL by 8.05 and 5.08 PQ points, respectively, suggesting that our framework makes more effective use of multimodal information even under severe image degradation. Furthermore, RAMP3D consistently outperforms the multimodal IAL variants equipped with DCGL, xMUDA, and UniDSeg across all three target domains.

Cross-dataset Domain Generalization. RAMP3D also improves generalization from nuScenes to SemanticKITTI and Waymo, which introduce shifts in sensor configurations and scene distributions. Compared with multimodal IAL, it improves PQ by 20.28 and 23.65 points on SemanticKITTI and Waymo, respectively. Image coverage further differs across datasets: nuScenes provides surround-

view images, whereas SemanticKITTI provides only forward-facing views. Such differences may reduce the effectiveness of fusion learned under the source-domain camera coverage and contribute to multimodal IAL underperforming its LiDAR-only counterpart. RAMP3D surpasses both multimodal and LiDAR-only IAL on both target datasets, suggesting that our framework effectively generalizes across image coverage and sensor configuration shifts.

Panoptic Error Maps. Fig. 3 compares the panoptic error maps of RAMP3D, xMUDA, and UniDSeg on the three unseen intra-nuScenes target domains: nighttime, rainy, and Singapore. Point colors indicate different error types: **false positives** and **false negatives** reflect recognition errors, while **well-matched** and **mismatched** points within true-positive segments reveal segmentation quality. Across all three target domains, RAMP3D produces visibly cleaner error maps than xMUDA and UniDSeg, reducing recognition and segmentation errors for both thing instances and stuff regions. These results demonstrate its robust one-to-many domain generalization to multiple unseen domains.

4.4 Ablation Studies

Effectiveness of ReMA. Table 2 evaluates the contributions of geometric guidance and gating in ReMA. All variants use the same frozen DINOv2 backbone.

Without gating, adding 3D feature injection (5th row) to 2D injection (3rd row) improves PQ across all three target domains, with the largest gain of 2.10 PQ points on nighttime. This supports the value of geometric guidance for visual adaptation under challenging illumination conditions.

Table 2: ReMA ablation with a frozen DINOv2 backbone. A+L: augmentation and auxiliary loss; Inject.: injected modality features; G/L: global/local gating.

A+L	Inject.	Gate	Source	Night	Rainy	SG
✗	–	–	77.38	42.08	71.99	64.69
✓	–	–	80.49	42.75	72.42	64.65
✓	2D	–	79.48	44.01	74.51	66.41
✓	3D	–	80.05	46.75	75.42	65.23
✓	2D+3D	–	80.32	46.11	75.62	66.88
✓	2D+3D	G	80.10	46.76	76.07	67.51
✓	2D+3D	L	**80.87**	48.51	76.59	68.36
✓	2D+3D	G+L	80.65	**49.34**	**76.83**	**68.85**

Gating provides further improvements beyond joint feature injection. Compared with no gating (5th row), combining global and local gating (last row) improves PQ by 3.23, 1.21, and 1.97 points on nighttime, rainy, and Singapore, respectively. Local gating consistently outperforms global gating, including a gain of 1.75 PQ points on nighttime, supporting the benefit of finer-grained adjustments. Combining global and local gating further improves over local gating alone on all three targets, most notably on nighttime, supporting their joint use for multimodal adaptation.

Effectiveness of ReCal. Table 3 evaluates ReCal on top of the full ReMA model. Compared with the model without ReCal (1st row), adding stuff calibration (2nd row) improves PQ across all three target domains, with the largest gain of 1.60 PQ points on nighttime. Adding thing re-scoring on top of stuff calibration (last row) provides further gains of 1.12 PQ points on nighttime. The combined configuration improves over the model without ReCal. These results support combining stuff mask calibration with thing candidate re-scoring, with the largest overall benefit observed under nighttime conditions.

Table 3: Ablation of ReCal modules. Thing: candidate re-scoring; Stuff: mask calibration.

Thing	Stuff	Source	Night	Rainy	SG
✗	✗	**80.65**	49.34	76.83	68.85
✗	✓	80.59	50.94	77.05	68.94
✓	✗	79.98	51.02	77.60	68.50
✓	✓	80.12	**52.06**	**77.65**	**69.02**

5 Conclusion

We presented RAMP3D, the first framework for domain-generalized multimodal 3D panoptic segmentation that addresses reliability at both the representation and prediction levels. ReMA estimates modality reliability to regulate image- and LiDAR-derived residual updates within a frozen VFM. ReCal uses a shared within-scene road reference to calibrate stuff masks and re-score thing candidates based on estimated mask quality, addressing their distinct prediction requirements. Extensive experiments across illumination, weather, geographic, and sensor shifts demonstrate consistent improvements over strong baselines, including LiDAR-based DG and multimodal UDA methods adapted to the source-only setting.

AI USE STATEMENT

We used generative AI tools solely for writing assistance, including grammar correction and sentence-level rephrasing of text originally drafted by the authors. We also used an AI coding assistant for routine implementation tasks, such as plotting scripts and data-loading utilities. Generative AI was not used to formulate research ideas, design experiments, or interpret experimental results. All AI-assisted text was reviewed by the authors, and all AI-assisted code was verified against the reported results. The authors take full responsibility for the final content of this work, including its text, claims, results, and associated artifacts.

REFERENCES

J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss. Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset. *The International Journal on Robotics Research*, 40(8-9):959–967, 2021. doi: 10.1177/02783649211006735.

Borna Besic, Nikhil Gosala, Daniele Cattaneo, and Abhinav Valada. Unsupervised domain adaptation for lidar panoptic segmentation. *IEEE Robotics and Automation Letters*, 7(2):3404–3411, 2022. ISSN 2377-3766, 2377-3774. doi: 10.1109/LRA.2022.3147326. URL https://ieeexplore.ieee.org/document/9698989/.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Exploiting the complementarity of 2d and 3d networks to address domain-shift in 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 98–109, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/WAD/html/Cardace_Exploiting_the_Complementarity_of_2D_and_3D_Networks_To_Address_CVPRW_2023_paper.html.

Siyu Chen, Ting Han, Changshe Zhang, Xin Luo, Meiliu Wu, Guorong Cai, and Jinhe Su. Stronger, steadier & superior: Geometric consistency in depth vfm forges domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Jeong Joon Park, Heng Yang, and Z. Morley Mao. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In *International Conference on Learning Representations (ICLR)*, 2025.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yi Feng, Junwu E, Zizhan Guo, Yu Ma, Hanli Wang, and Rui Fan. An instance-centric panoptic occupancy prediction benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yinjuan Gu, Yuming Huang, Chengzhong Xu, and Hui Kong. Maskrange: A mask-classification model for range-view based lidar segmentation. *ArXiv*, abs/2206.12073, 2022. URL https://api.semanticscholar.org/CorpusID:250048842.

Pei He, Lingling Li, Licheng Jiao, Ronghua Shang, Fang Liu, Shuang Wang, Xu Liu, and Wenping Ma. Domain-aware category-level geometry learning segmentation for 3d point clouds. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. LiDAR-based Panoptic Segmentation via Dynamic Shifting Network. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13085–13094, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.01289. URL https://ieeexplore.ieee.org/document/9577774/.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3D semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Hyeonseong Kim, Yoonsu Kang, Changgyoon Oh, and Kuk-Jin Yoon. Single domain generalization for lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Fan Li, Xuan Wang, Xuanbin Wang, Zhaoxiang Zhang, and Yuelei Xu. No object is an island: Enhancing 3d semantic segmentation generalization with diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11799–11808, 2022. URL https://api.semanticscholar.org/CorpusID:248811224.

Miaoyu Li, Yachao Zhang, Xu Ma, Yanyun Qu, and Yun Fu. Bev-dg: Cross-modal learning under bird's-eye view for domain generalization of 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Xiao Li, Hui Li, Xiangzhen Kong, Yuang Ji, Zhiyu Liu, and Hao Liu. Panoptic-vsnet: Visual-semantic prior knowledge-driven multimodal 3d panoptic segmentation. *Pattern Recognition*, 176: 113239, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113239. URL https://www.sciencedirect.com/science/article/pii/S0031320326002049.

R. Marcuzzi, L. Nunes, L. Wiesmann, J. Behley, and C. Stachniss. Mask-Based Panoptic LiDAR Segmentation for Autonomous Driving. *IEEE Robotics and Automation Letters*, 8(2):1141–1148, 2023. doi: 10.1109/LRA.2023.3236568. URL https://www.ipb.uni-bonn.de/wp-content/papercite-data/pdf/marcuzzi2023ral.pdf.

Rohit Mohan, Florian Drews, Yakov Miron, Daniele Cattaneo, and Abhinav Valada. Up-fuse: Uncertainty-guided lidar-camera fusion for 3d panoptic segmentation, 2026. URL https://arxiv.org/abs/2602.19349.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Marin Oršić and Siniša Šegvić. Efficient semantic segmentation with pyramidal fusion. *Pattern Recognition*, 110:107611, 2021. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107611. URL https://www.sciencedirect.com/science/article/pii/S0031320320304143.

Yining Pan, Qiongjie Cui, Xulei Yang, and Na Zhao. How do images align and complement LiDAR? Towards a harmonized multi-modal 3D panoptic segmentation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47615–47631. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pan25c.html.

Yining Pan, Shijie Li, Yuchen Wu, Xulei Yang, and Na Zhao. Panda: Unsupervised domain adaptation for multimodal 3d panoptic segmentation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ryan Razani, Ran Cheng, Enxu Li, Ehsan Moeen Taghavi, Yuan Ren, and Bingbing Liu. Gp-s3net: Graph-based panoptic sparse semantic segmentation network. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16056–16065, 2021. URL https://api.semanticscholar.org/CorpusID:237213370.

Cristiano Saltori, Aljoša Ošep, Elisa Ricci, and Laura Leal-Taixé. Walking your lidog: A journey through multiple domains for lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Kshitij Sirohi, Rohit Mohan, Daniel Buscher, Wolfram Burgard, and Abhinav Valada. Efficientlps: Efficient lidar panoptic segmentation. *IEEE Transactions on Robotics*, 38:1894–1914, 2021. URL https://api.semanticscholar.org/CorpusID:231933783.

Hamin Song, Jieun Cho, Jinsu Ha, Jaehyun Park, and Kichun Jo. Panoptic-fusionnet: Camera-lidar fusion-based point cloud panoptic segmentation for autonomous driving. *Expert Syst. Appl.*, 251(C), October 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.123950. URL https://doi.org/10.1016/j.eswa.2024.123950.

Nicolai Steinke, Daniel Goehring, and Raúl Rojas. Groundgrid: Lidar point cloud ground segmentation and terrain estimation. *IEEE Robotics and Automation Letters*, 9(1):420–426, 2024. doi: 10.1109/LRA.2023.3333233.

Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Thang Vu, Kookhoi Kim, Tung M. Luu, Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28619–28630, June 2024.

Yao Wu, Mingwei Xing, Yachao Zhang, Xiaotong Luo, Yuan Xie, and Yanyun Qu. Unidseg: Unified cross-domain 3d semantic segmentation via visual foundation models prior. In *Proceedings of the 38th International Conference on Neural In- formation Processing Systems*, 2024.

Yuchen Wu, Kun Wang, Yining Pan, and Na Zhao. Ccf: Complementary collaborative fusion for domain generalized multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026. doi: 10.48550/arXiv.2603.23276.

Zeqi Xiao, Wenwei Zhang, Tai Wang, Chen Change Loy, Dahua Lin, and Jiangmiao Pang. Position-Guided Point Cloud Panoptic Segmentation Transformer. *International Journal of Computer Vision*, 133(1):275–290, January 2025. ISSN 1573-1405. doi: 10.1007/s11263-024-02162-z. URL https://doi.org/10.1007/s11263-024-02162-z.

Shuangjie Xu, Rui Wan, Maosheng Ye, Xiaoyi Zou, and Tongyi Cao. Sparse cross-scale attention network for efficient lidar panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2920–2928, 2022.

Xiang Xu, Lingdong Kong, Hui Shuai, and Qingshan Liu. Frnet: Frustum-range networks for scalable lidar segmentation. *IEEE Transactions on Image Processing*, 34:2173–2186, 2025.

Yixuan Xu, Hamidreza Fazlali, Yuan Ren, and Bingbing Liu. Aop-net: All-in-one perception network for lidar-based joint 3d object detection and panoptic segmentation. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–7, 2023. doi: 10.1109/IV55152.2023.10186743.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *International Conference on Learning Representations (ICLR)*, 2024.

Dongqiangzi Ye, Zixiang Zhou, Weijia Chen, Yufei Xie, Yu Wang, Panqu Wang, and Hassan Foroosh. Lidarmultinet: towards a unified multi-task network for lidar perception. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i3.25429. URL https://doi.org/10.1609/aaai.v37i3.25429.

Hongqi Yu, Sixian Chan, Xiaolong Zhou, and Xiaoqin Zhang. Sgformer: semantic-geometry fusion transformer for multi-modal 3d panoptic segmentation. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i9.33042. URL https://doi.org/10.1609/aaai.v39i9.33042.

Qi Zang, Dong Zhao, Nan Pu, Wenjing Li, Zhun Zhong, and Meng Wang. Geco: Geometry-consistent regularization for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Hongjia Zhai, Qi Zhang, Xiaokun Pan, Xiyu Zhang, Yitong Dong, Huaqi Zhang, Dan Xu, and Guofeng Zhang. Onlinepg: Online open-vocabulary panoptic mapping with 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026. URL https://arxiv.org/abs/2603.18510.

Zhiwei Zhang, Zhizhong Zhang, Qian Yu, Ran Yi, Yuan Xie, and Lizhuang Ma. Lidar-camera panoptic segmentation via geometry-consistent and semantic-aware alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3662–3671, October 2023.

Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-PolarNet: Proposal-free LiDAR Point Cloud Panoptic Segmentation . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13189–13198, June 2021. doi: 10.1109/CVPR46437.2021.01299. URL https://doi.ieeecomputersociety.org/10.1109/CVPR46437.2021.01299.

Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. *arXiv preprint arXiv:2011.10033*, 2020.

A Cross-Domain Benchmark Setting Details

A.1 Intra-Dataset Domain Shifts

We evaluate intra-dataset domain generalization on nuScenes under a one-to-many setting. Following the preprocessing pipeline of closed-domain 3D panoptic segmentation Zhou et al. (2021); Pan et al. (2025); Xiao et al. (2025), we use all available camera views and LiDAR points for full-scene evaluation. Using the official training/validation split and scene metadata, we construct a single source domain comprising Boston scenes captured under sunny daytime conditions. Only source-domain samples from the official training split are used for training. Here, *sunny* denotes scenes without a rainy annotation. We evaluate the trained model on three unseen target domains drawn from the official validation split: *Singapore*, *rainy*, and *nighttime*. Each target domain is defined by its corresponding attribute, with the remaining attributes unconstrained. The target subsets may therefore overlap; for example, a rainy nighttime scene in Singapore belongs to all three subsets. No target-domain samples are used during training.

We use the official class mapping with 10 *thing* and 6 *stuff* classes. For target domains in which some classes are absent, we average the evaluation metrics over the remaining classes present in each target set. Dataset statistics are summarized in Table 4, and representative LiDAR and image samples are shown in Fig. 4. The three target domains probe different, potentially co-occurring shifts: **Geographic Shift: Singapore.** Compared with the Boston source domain, Singapore introduces differences in architectural styles, vegetation, road layouts, and scene composition. These changes affect both geometric and visual representations, testing generalization across geographic environments. **Weather Shift: Rainy.** Rain can affect both modalities. LiDAR observations may exhibit spurious returns, missing points, and reflectance variations, while images may suffer from raindrops, reduced visibility, and lower contrast. These effects alter the quality and complementarity of geometric and visual cues, particularly for small or distant objects. **Illumination Shift: Nighttime.** Nighttime primarily challenges the image modality through low brightness, increased noise, and weakened texture and boundary cues. This setting tests whether multimodal models can regulate their reliance on visual information when its reliability decreases.

Table 4: One-to-many DG settings, sample counts, and label spaces. Only source samples are used for training. Counts refer to LiDAR frames.

Role	Dataset / domain	# Train	# Eval.	Classes	Excluded categories
Intra-dataset: one source, three targets					
Source	nuScenes: Boston, sunny, daytime	10,635	2,122	$\mathcal{C}_{16}$	–
Target	nuScenes: Singapore	–	2,929	$\mathcal{C}_{16}$	trailer
Target	nuScenes: rainy	–	1,088	$\mathcal{C}_{16}$	–
Target	nuScenes: nighttime	–	602	$\mathcal{C}_{16}$	bus, construction_vehicle, trailer
Cross-dataset: one source, two targets					
Source	nuScenes	28,130	6,019	$\mathcal{C}_6$	
Target	SemanticKITTI	–	1,101	$\mathcal{C}_6$	refer to Table 5
Target	Waymo	–	5,976	$\mathcal{C}_6$	

$\mathcal{C}_{16}$: **Thing (10):** barrier, bicycle, bus, car, construction_vehicle, motorcycle, pedestrian, traffic_cone, trailer, truck. **Stuff (6):** driveable_surface, other_flat, sidewalk, terrain, manmade, vegetation.

$\mathcal{C}_6$: **Thing (1):** vehicle. **Stuff (5):** driveable_surface, sidewalk, terrain, manmade, vegetation.

A.2 Cross-Dataset Domain Shift

We evaluate cross-dataset domain generalization in a one-to-many setting, using nuScenes as the single source domain and SemanticKITTI and Waymo as two unseen target domains. The model is trained once on the full official nuScenes training split and evaluated on both SemanticKITTI sequence 07 and the Waymo validation split, without accessing either target dataset during training.

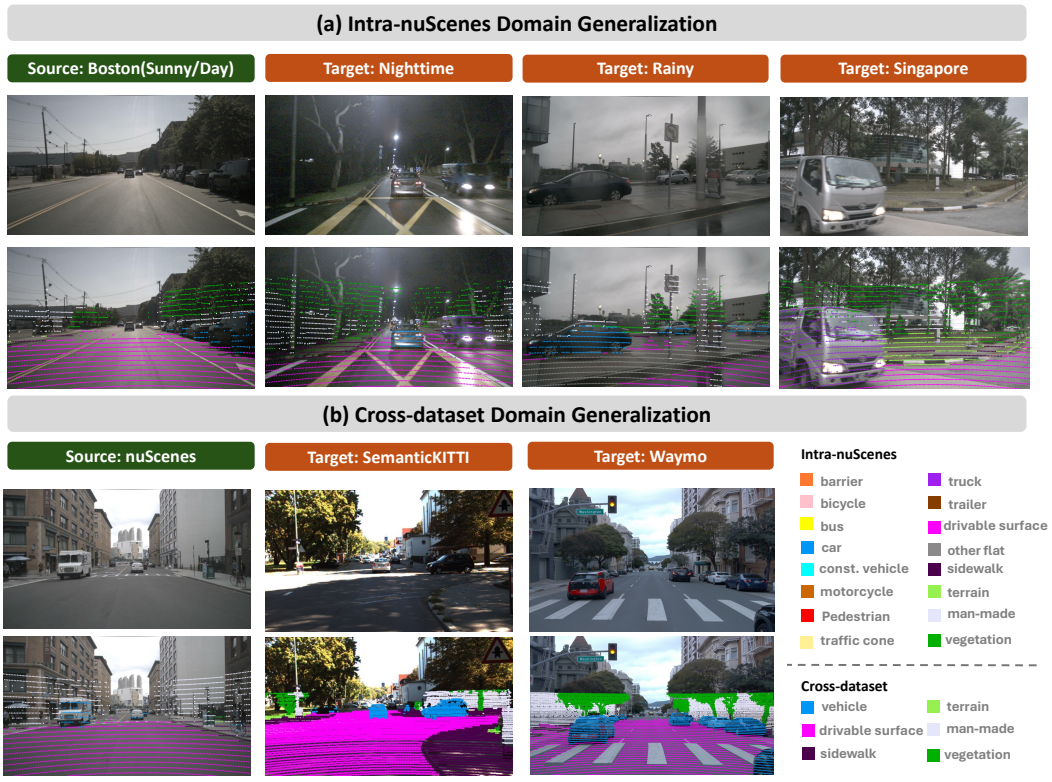


Figure 4: Illustration of cross-domain benchmarks. To highlight domain discrepancies, LiDAR points are projected onto front-view images and color-coded by semantic classes. Best viewed in color.

To reconcile the different class taxonomies, we map annotations into a shared label space containing one *thing* class, *vehicle*, and five *stuff* classes: *driveable surface*, *sidewalk*, *terrain*, *manmade*, and *vegetation*. Categories without consistent cross-dataset correspondence are assigned to *ignore*. The exact mappings are provided in Table 5.

The two target datasets introduce compound shifts in sensor configuration, camera coverage, and scene distribution. SemanticKITTI differs from nuScenes in LiDAR sampling characteristics and provides two front-view camera streams, compared with the six surround-view cameras in nuScenes. For Waymo, we use the first return of the TOP LiDAR together with five camera views, preserving the native range-image coordinates before converting the point cloud to the common range-view representation. These sensor differences, together with changes in the captured environments, test whether a single source-trained model can generalize across multiple unseen domains without target-specific adaptation.

B Additional Qualitative Results

Fig. 5 presents extended panoptic error map visualizations, complementing Fig. 3 in the main paper. Denser colored regions indicate more prediction errors. The highlighted regions demonstrate improvements on both thing and stuff categories. For thing instances, RAMP3D reduces **false positives** and **false negatives**, while recovering more complete and coherent object instances. For stuff regions, it produces clearer semantic boundaries and fewer **mismatched** points. These improvements hold under nighttime illumination, rainy weather, and the geographic shift to Singapore, demonstrating robust one-to-many domain generalization.

Table 5: Class mapping from SemanticKITTI, nuScenes and Waymo to the six shared classes (vehicle, driveable_surface, sidewalk, terrain, manmade, vegetation) used in the N→K/W setting.

SemanticKITTI	**mapped classes**	**nuScenes**	**mapped classes**	**Waymo**	**mapped classes**
unlabeled	ignore	ignore	ignore	unknown	ignore
outlier	ignore	barrier	ignore	car	vehicle
car	vehicle	bicycle	vehicle	truck	vehicle
bicycle	vehicle	bus	vehicle	bus	vehicle
bus	ignore	car	vehicle	other_vehicle	vehicle
motorcycle	vehicle	construction_vehicle	vehicle	motorcyclist	vehicle
on-rails	ignore	motorcycle	vehicle	bicyclist	vehicle
truck	vehicle	pedestrian	ignore	pedestrian	ignore
other-vehicle	ignore	traffic_cone	ignore	sign	manmade
person	ignore	trailer	vehicle	traffic_light	manmade
bicyclist	vehicle	truck	vehicle	pole	manmade
motorcyclist	vehicle	driveable_surface	driveable_surface	construction_cone	ignore
road	driveable_surface	other_flat	ignore	bicycle	vehicle
parking	driveable_surface	sidewalk	sidewalk	motorcycle	vehicle
sidewalk	sidewalk	terrain	terrain	building	manmade
other-ground	ignore	manmade	manmade	vegetation	vegetation
building	manmade	vegetation	vegetation	tree_trunk	vegetation
fence	manmade			curb	ignore
other-structure	ignore			road	driveable_surface
lane-marking	driveable_surface			lane_marker	driveable_surface
vegetation	vegetation			other_ground	ignore
trunk	vegetation			walkable	terrain
terrain	terrain			sidewalk	sidewalk
pole	manmade				
traffic-sign	manmade				
other-object	manmade				
moving-car	vehicle				
moving-bicyclist	vehicle				
moving-person	ignore				
moving-motorcyclist	vehicle				
moving-on-rails	ignore				
moving-bus	ignore				
moving-truck	vehicle				
moving-other-vehicle	ignore				



Figure 5: Additional panoptic error map visualizations on the nighttime, rainy, and Singapore target domains under the one-to-many domain generalization setting. We compare RAMP3D with xMUDA and UniDSeg. Point colors indicate different error types: **false positives** and **false negatives** reflect recognition errors, while **well-matched** and **mismatched** points within true-positive segments indicate segmentation quality. Camera images and semantic-colored ground-truth point clouds are provided for reference. Stuff-region boundaries are highlighted with **scarlet** boxes, whereas thing-instance segmentation results are highlighted with **gold** boxes. Best viewed in color.

Table 6: Model complexity and inference latency.

Model	**Total params. (M)**	**Trainable (M)**	**Latency (ms)**
Baseline	64.87	64.87	180.3
Frozen DINOv2	145.59	59.41	278.5
+ ReMA	152.10	65.91	291.5
+ ReMA + ReCal	152.21	66.02	298.2

Table 7: Performance and parameter counts across different 3D and 2D backbones.

3D Backbone	2D Backbone	Source	Night	Rainy	SG	Params (M)	Trainable (M)
FRNet	SwiftNet	76.25	41.62	68.75	60.24	64.9	64.9
FRNet	DINO-B	80.12	52.06	77.65	69.02	152.2	66.0
FRNet	DINO-L	82.06	53.22	78.18	70.74	373.8	69.9
FRNet	CLIP-B	80.29	48.42	75.96	67.28	151.8	66.0
FRNet	CLIP-L	80.47	48.74	76.28	67.44	373.1	69.9
Cylinder3D	DINO-B	83.74	52.83	78.17	69.17	173.1	86.9

C Model Complexity and Inference Efficiency

Table 6 shows the model parameters and inference latency of our proposed framework. The baseline is IAL with an FRNet LiDAR encoder and a SwiftNet image encoder; the frozen-DINOv2 variant replaces SwiftNet with a frozen DINOv2 ViT-B/14. Our full model contains 66.02M trainable parameters, only 1.15M (1.8%) more than the baseline. Although latency increases from 180.3 to 298.2 ms, most of this increase is already incurred by the frozen-DINOv2 configuration. Relative to this VFM baseline, ReMA and ReCal jointly add only 6.62M total parameters (4.5%) and 19.7 ms of latency (7.1%). ReCal alone contributes 0.11M parameters and 6.7 ms. The complete RAMP3D framework improves PQ by up to 10.44% over the original baseline on the nighttime domain (in Table 1), achieving improved generalization with modest additional overhead from the proposed modules. We measure inference latency on an NVIDIA RTX 5090 GPU with a batch size of 1.

D Framework Generality

We further evaluate the compatibility of our framework with different 3D representations and visual backbones. As shown in Table 7, our method can be used with both the range-view-based FRNet Xu et al. (2025) and the voxel-based Cylinder3D Zhu et al. (2020), indicating that it is not tied to a specific 3D representation. With FRNet, replacing the conventional SwiftNet Oršić & Šegvić (2021) image encoder with different vision foundation models consistently improves performance on both the source and unseen domains. In particular, DINO-B and DINO-L substantially improve nighttime, rainy, and Singapore performance, while CLIP-B and CLIP-L also provide consistent gains. Increasing the foundation-model capacity from the base to large variant further improves performance, with DINO-L achieving the strongest results among the evaluated visual backbones. These results demonstrate that our framework is broadly compatible with different 3D representations and visual foundation models, rather than relying on a particular backbone configuration.

E Broader Impact and Limitations

Broader Impact. To the best of our knowledge, this work is the first to systematically study domain generalization for multimodal 3D panoptic segmentation, establishing a new setting for robust multimodal scene understanding under unseen domains. Unlike domain adaptation, domain generalization requires no access to target-domain data, making it particularly relevant to industrial deployment, where autonomous systems must continuously operate across different cities, weather conditions, and sensor environments without repeated data collection, annotation, or model adaptation. Moreover, 3D panoptic segmentation provides a unified understanding of both semantic regions and

individual object instances, offering richer scene representations than semantic segmentation alone and supporting downstream tasks such as autonomous navigation, planning, and decision-making. We hope this work can facilitate the development of more scalable and robust multimodal perception systems for real-world autonomous driving.

Potential Limitations. Our framework relies on multimodal inputs from both LiDAR and cameras, introducing additional hardware requirements for data acquisition, deployment, and inference compared with single-modality systems. In addition, ReMA builds on pretrained visual foundation models, and its performance can be affected by the quality of VFM representations, whose errors may propagate under highly ambiguous scenes. Reducing multimodal sensing and computational overhead, as well as improving robustness to imperfect foundation-model representations, are important directions for future work.